*Q. J. R. Meteorol. Soc.* (2000), **126**, pp. 649–667

# Skill and relative economic value of the ECMWF ensemble prediction system

By D. S. RICHARDSON*
*European Centre for Medium-Range Weather Forecasts, UK*

(Received 8 October 1998; revised 22 June 1999)

## SUMMARY

The economic value of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational ensemble prediction system (EPS) is assessed relative to the value of a perfect deterministic forecast. The EPS has substantial relative value throughout the medium range. Probability forecasts derived from the EPS are of greater benefit than a deterministic forecast produced by the same model. Indeed, for many users, the probability forecasts have more value than a shorter-range deterministic forecast. Based on the measures used here, the additional information in the EPS (reflecting the uncertainty in the initial conditions) provides a benefit to users equivalent to many years' development of the forecast model and assimilation system.

The impact of ensemble size on forecast value is considered. The difference in performance between ensembles with 10 and with 50 members may appear relatively small, based on standard skill measures, yet the larger ensembles have substantial benefit to a range of users. Further increases in ensemble size may be expected to provide additional value.

KEYWORDS: Numerical weather prediction Probability forecasts

## 1. INTRODUCTION

Ensemble predictions have been produced regularly at the European Centre for Medium-Range Weather Forecasts (ECMWF) since December 1992 (Palmer *et al.* 1993; Molteni *et al.* 1996). The current operational Ensemble Prediction System (EPS) comprises a control forecast initialized from the operational analysis, plus 50 additional integrations initialized from perturbations to the control analysis (Buizza *et al.* 1998). All forecasts are made with the operational ECMWF model, but run at lower horizontal resolution ($T_L159$) than the single high-resolution deterministic forecast ($T_L319$). The EPS complements the deterministic forecast by the provision of information about the probability distribution of future weather, based on uncertainty in the initial analysis.

The performance of the EPS is routinely monitored using a range of verification measures. These assessments demonstrate that the EPS is a skilful prediction system. They also demonstrate the improvement of the enhanced EPS introduced in December 1996 (Buizza *et al.* 1998). However, they do not explicitly address the question which is perhaps of most concern to potential users, viz. 'Is the EPS worth paying for?'.

Providing an answer to such a question is not straightforward. To obtain economic benefit from a forecast, a potential user must have alternative courses of action available, whose consequences will depend on the weather that occurs. If, by using forecasts, the user decides on actions he would not otherwise take, and benefits economically from these alternative actions, then the forecasts have been of value to the user (Murphy 1994). Thus, a proper evaluation of the benefits of a forecast system to a particular user will involve not only the intrinsic skill of the forecasts, but also detailed knowledge of the user's decision-making processes and the exact nature of the weather-sensitivity of that user's operation.

Several authors have investigated the relationship between skill and economic value of weather forecasts, either using specific examples (Roebber and Bosart 1996), or in general terms (Murphy and Ehrendorfer 1987; Wilks and Hamill 1995). Katz and Murphy (1997a) provided a comprehensive account of such research, and an extensive collection of references. These studies showed that the relationship between skill and

* Corresponding address: European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

TABLE 1. COST AND LOSS FOR DIFFERENT OUTCOMES

| | | Occurs | |
|---|---|---|---|
| | | No | Yes |
| Take action | No | 0 | $L$ |
| | Yes | $C$ | $C + L - L_1$ |

value is complex. Reliance on measures of skill alone may give a misleading impression of forecast value. Decisions on system configuration (for example, choice of optimal ensemble-size), may be different if the effect on economic value is considered rather than just the overall skill. Consequently, there are two reasons for studying the economic value of the EPS: to evaluate the potential benefit to users and to assess the impact of possible system changes in a context relevant to users.

The present paper uses a simple decision-analytic model to examine the economic value of the EPS relative to that of a (hypothetical) perfect deterministic forecast. As examples of the operational EPS products, let us consider predicted probabilities of 850 hPa temperature-anomaly over Europe exceeding certain thresholds. The decision model (introduced in section 2) is applied to the deterministic control-forecast (section 3) and to the EPS probability-forecasts (section 4) for these predictions for January and February 1998. The benefit of the EPS over the deterministic forecast is assessed in section 5. In section 6, the effect of ensemble size on forecast value is studied, using EPS precipitation-forecast data for winter 1996–97. Conclusions are drawn in section 7.

The decision-analytic model discussed in the present paper has also been applied to seasonal forecast ensembles: the results have been discussed by Palmer *et al.* (2000).

## 2. THE COST–LOSS RATIO DECISION MODEL

The economic value of weather forecasts has often been discussed in terms of so-called decision-analytic models (e.g., Murphy 1977; Katz and Murphy 1997b). A decision maker has a number of courses of action to choose from, and the choice is to some extent influenced by the forecast. Each action has an associated cost and leads to an economic benefit or loss depending on the weather outcome. The task is to choose the appropriate action which will minimize the expected loss. To emphasize the main points of the decision framework, let us consider here the simplest decision model, known as the static cost–loss model.

Consider a decision maker who must choose either to take action or to do nothing, the choice depending exclusively on his belief that a given weather event $X$ will occur or not. If the event occurs and no action has been taken then the decision maker incurs a loss $L$. Taking action incurs a cost $C$ irrespective of the outcome; if the event occurs, the action prevents part of the loss ($L_1$). For example, $X$ might be the occurrence of ice on roads and the action 'to grit the roads'; $C$ would be the cost of the gritting procedure and $L$ would be the economic loss as a result of traffic delays and accidents on icy roads. The expense associated with each combination of action and occurrence of $X$ is shown in Table 1 (the expense matrix).

The decision maker wishes to pursue the strategy which will minimize his expense over a large number of cases. If only climatological information is available there are two options: either always take protective action or never protect. Always taking action incurs a cost $C$ on each occasion and a loss $L - L_1$, on that fraction $\bar{o}$ of occasions when the event occurs, hence the average expense is $C + \bar{o}(L - L_1)$. If action is never taken,

TABLE 2. CONTINGENCY TABLE FOR FORECAST AND OCCURRENCE OF BINARY EVENT

| | | Observed | | |
|---|---|---|---|---|
| | | No | Yes | |
| Forecast | No | $a$ | $b$ | $a+b$ |
| | Yes | $c$ | $d$ | $c+d$ |
| | | $a+c=1-\bar{o}$ | $b+d=\bar{o}$ | $a+b+c+d=1$ |

the loss $L$ occurs only when the event occurs, and the average expense is $\bar{o}L$. Thus in the absence of information other than climatology, the optimal course of action is always to act if $C+\bar{o}(L-L_1)<\bar{o}L$, i.e. $C<\bar{o}L_1$, and never act otherwise; the expected expense is then

$$E_{\text{climate}}=\min\{C+\bar{o}(L-L_1),\bar{o}L\}. \tag{1}$$

The provision of additional information in the form of forecasts may allow the decision maker to revise his strategy and reduce his expected expense. The reduction in expense is a measure of the value of the forecasts to the decision maker.

Given perfect knowledge of the future weather, the decision maker would need to take action only when the event was going to occur. The expected expense would then be

$$E_{\text{perfect}}=\bar{o}(C+L-L_1). \tag{2}$$

Let us define the relative value $V$ of a forecast system as the reduction in expense as a proportion of that which would be achieved by a perfect forecast:

$$V=\frac{E_{\text{climate}}-E_{\text{forecast}}}{E_{\text{climate}}-E_{\text{perfect}}}. \tag{3}$$

Consequently, maximum relative value $V=1$ will be obtained from a perfect forecast system, and $V=0$ for a climate forecast. If $V>0$ then the user will benefit from the system. This definition gives an absolute upper bound to $V$ and is a convenient reference level for the user: if a perfect knowledge of the future weather will save the user an amount $S$ (over the use of purely climatological information) then the EPS with relative value $V$ will save the user $100V\%$ of $S$.

## 3. SKILL AND RELATIVE VALUE FOR A DETERMINISTIC FORECAST-SYSTEM

Consider first a deterministic forecast-system, that is each forecast is a simple statement either that a weather event $X$ will occur or that it will not occur. The performance of the system over a period of time can be summarized in a contingency table which shows the fraction of correct and incorrect forecasts of a weather event occurring or not occurring (Table 2).

The hit rate $H$ is defined as the fraction of occurrences of the event which were correctly forecast, while the false-alarm rate $F$ is the fraction of non-occurrences for which the event was (incorrectly) forecast.

$$H=d/(b+d)=d/\bar{o} \tag{4}$$

$$F=c/(a+c)=c/(1-\bar{o}). \tag{5}$$

Note that both $H$ and $F$ are expressed in terms of the observed relative frequency $\bar{o}$ of the event; it is assumed that $\bar{o}>0$, i.e. that the event does occur in the sample.

TABLE 3. SKILL OF THE EPS CONTROL 6-DAY FORECASTS OF 850 hPa TEMPERATURE ANOMALIES OVER EUROPE

| | January and February 1998 | | | |
|---|---|---|---|---|
| Event | $F$ | $H$ | $KS$ | $\bar{o}$ |
| $T < -8$ K | 0.039 | 0.445 | 0.406 | 0.058 |
| $T < -4$ K | 0.144 | 0.611 | 0.468 | 0.228 |
| $T > +4$ K | 0.091 | 0.548 | 0.457 | 0.179 |
| $T > +8$ K | 0.027 | 0.393 | 0.367 | 0.043 |

From Table 2 and the expense matrix (Table 1), the expected expense $E$ for the forecast system is

$$E = bL + cC + d(C + L - L_1). \tag{6}$$

This may be written in terms of $H$ and $F$ using Eqs. (4) and (5) as

$$E = F(1 - \bar{o})C - H\bar{o}(L_1 - C) + \bar{o}L. \tag{7}$$

Substituting from Eqs. (1), (2), and (7) into Eq. (3), the relative value of the forecast system is

$$V = \frac{\min(\alpha, \bar{o}) - F\alpha(1 - \bar{o}) + H\bar{o}(1 - \alpha) - \bar{o}}{\min(\alpha, \bar{o}) - \bar{o}\alpha}, \tag{8}$$

where $\alpha = C/L_1$, the cost of taking action expressed as a fraction of potential loss protected by the action, is known as the 'cost–loss' ratio. As a result, the relative value of a particular forecast-system depends on parameters $\alpha$ and $\bar{o}$, which are external to the system, and $H$ and $F$ which are model-dependent.

Various measures of forecast skill may be derived from Table 2 (Wilks 1995; Stanski *et al.* 1989). Here, let us use the Kuipers score (*KS*) which has the desirable characteristic of equitability (Gandin and Murphy 1992), in that random or constant forecasts will score zero and perfect forecasts will have a score of 1. In the notation of Table 2, this may be written as

$$KS = \frac{ad - bc}{(a + c)(b + d)} = H - F. \tag{9}$$

To illustrate the performance of a deterministic forecast-system, let us consider the control integration (the forecast from the unperturbed analysis). Table 3 shows results for the prediction of day-6 850 hPa temperature anomalies exceeding certain thresholds over Europe in January and February 1998. The control forecast has substantial skill for all thresholds. Forecasts for the smaller thresholds are more skilful than forecasts for the more extreme events, and positive anomalies appear more difficult to predict than negative anomalies. However, the question for potential users is 'How does this skill translate to economic value of a forecast?'.

For a given weather event and forecast system, $\bar{o}$, $H$ and $F$ are given, so that the relative economic value $V$ of the forecast system depends only on $\alpha$. Figure 1 shows $V$ as a function of $\alpha$ for the forecasts of the four events. Although, according to the scores in Table 3, the model is skilful, the usefulness to a decision maker depends greatly on his particular cost–loss ratio. For $\alpha$ greater than about 0.6, none of the event forecasts are useful; for $0.1 < \alpha < 0.5$, forecasts of the ±4 K events are useful, and for $\alpha < 0.1$ only forecasts of large anomalies have value.

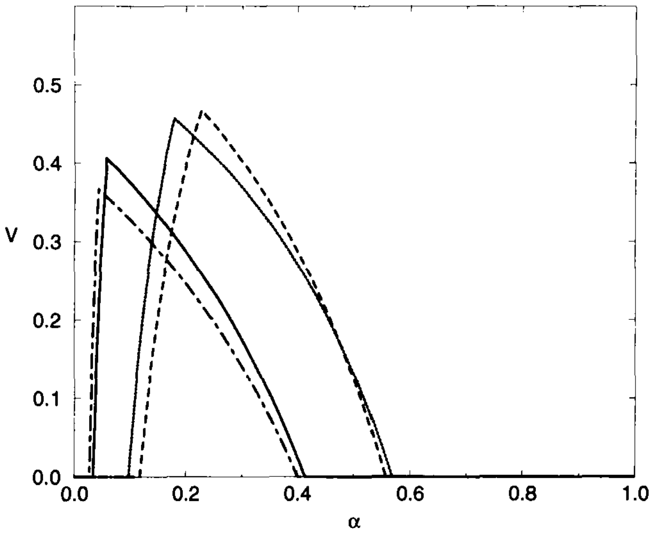

Figure 1. Ensemble Prediction System control deterministic forecasts of 850 hPa temperature anomalies $T$ exceeding four different thresholds over Europe at day 6 for January and February 1998: relative value $V$ plotted against cost–loss ratio $\alpha$ for events $T < -8$ K (solid line), $T < -4$ K (pecked line),$T > +4$ K (dotted line) and $T > +8$ K (dash-dotted line).

From Eq. (8), it is straightforward to determine the range of $\alpha$ for which $V$ is positive and also the maximum relative value. For $\alpha < \bar{o}$, Eq. (8) becomes

$$V = (1 - F) - \{\bar{o}/(1 - \bar{o})\}\{(1 - \alpha/\alpha\}(1 - H). \tag{10}$$

$V$ increases with $\alpha$ and is therefore greatest for $\alpha = \bar{o}$. It can be deduced (see Table 2 and the definitions of $H$ and $F$) that $V$ is greater than zero for $\alpha > b/(a + b)$; that is, for the forecasts to have value, the probability of the event $X$ occurring when it has not been forecast, $P(X = \text{Yes} \mid \text{forecast} = \text{No})$, must be lower than the user's $\alpha$. Similarly, for $\alpha > \bar{o}$,

$$V = H - \{(1 - \bar{o})/\bar{o}\}\{\alpha/1 - \alpha)\}F. \tag{11}$$

In this case, $V$ decreases with increasing $\alpha$, and is positive for $\alpha < d/(c + d)$; users will benefit from the forecasts if the probability of $X$ occurring when $X$ is forecast, $P(X = \text{Yes} \mid \text{forecast} = \text{Yes})$, exceeds $\alpha$.

So, maximum value always occurs for $\alpha = \bar{o}$; at this point, the expense of taking either of the climatological options (always protect or never protect) is the same: climatology does not help the decision maker and the forecast offers the greatest benefit. As the cost approaches the limits of 0 and 1, the climatological options become harder to beat—the great expense resulting from even occasional incorrect forecasts outweighs the small expenditure on the default action. The conditional probabilities of the event occurring, given the forecast, determine the range of $\alpha$ for which the forecast system is useful.

The maximum value is given by

$$V_{\text{max}} = H - F. \tag{12}$$

So, skill (Eq. (9)) is related to the usefulness of the forecasts: $KS$ is the maximum relative value that can be obtained from the system. Whether this potential maximum value will be achieved depends on the $\alpha$ of the user; the closer $\alpha$ is to $\overline{o}$ the higher will be the value. Note that this maximum value is independent of $\alpha$ and $\overline{o}$; if two systems predicting different events (with quite different values of $\overline{o}$) have the same *KS*, then the potential maximum value will be the same, but it will occur for different values of $\alpha$ (equal to the respective observed frequencies).

## 4. Probability forecasts

A user receiving forecasts expressed in terms of probabilities must decide on the probability threshold at which to take action. Should it be when the event is forecast with a probability of, say, 50%, or when the forecast is more certain (perhaps 80%)? Is there an optimum probability above which action should be taken?

In effect, this choice of a threshold probability, $p_t$, converts the probability forecast to a deterministic one. Consider those forecasts of a higher probability of the event as forecasts that the event will occur and those with lower probability as forecasts that the event will not occur. For a given value of $p_t$, the value of the system can then be determined in the same way as for a deterministic system. By varying $p_t$ from 0 to 1, a sequence of values for $H$ and $F$, and hence of $V$, can be derived; the user can then choose that value of $p_t$ which results in the largest value of $V$. Note that since $V$ also depends on $\overline{o}$ and $\alpha$, the appropriate value of $p_t$ will be different for different users and different weather events.

Probability forecasts of the temperature events considered in the previous section are produced using the EPS. The relative operating characteristic (ROC) (Mason 1982; Harvey *et al.* 1992) is a plot of $H$ against $F$ for a set of threshold probabilities $p_t$ between 0 and 1 (Fig. 2). The endpoints of the ROC (1, 1 and 0, 0) result from the baseline actions of always forecasting or never forecasting the event respectively. A perfect forecast system, with $H = 1$ and $F = 0$, would give a point at the top left-hand corner of the graph, so the closer the ROC is to the top left-hand corner the better. If a forecast system has no ability to discriminate occurrence of an event from non-occurrence, then $H$ and $F$ will always be equal and the ROC for the system would lie along the diagonal line $H = F$. The area A under the ROC is used as an index of the accuracy of the forecast system (Mason 1982; Buizza *et al.* 1998, 1999). A perfect system would have $A = 1.0$, while no-skill systems ($H = F$) would have $A = 0.5$. The area under the ROC of Fig. 2 is given in the caption. The area may be converted to a skill score relative to climatology and chance in the usual way (Stanski *et al.*, 1989) as

$$ASS = \frac{A_{\text{forecast}} - A_{\text{climate}}}{A_{\text{perfect}} - A_{\text{climate}}} = 2A - 1. \tag{13}$$

For each value of $p_t$, the corresponding values of $H$ and $F$ may be used to generate a value curve, just as in the deterministic case. The set of curves for $T > +4$ K is shown in Fig. 3. The EPS forecasts have value for most users, although the benefit varies substantially between users with different cost–loss ratios. The most important feature of Fig. 3 is that the relative value depends crucially on the appropriate choice of $p_t$. Users with small $\alpha$, i.e. relatively large potential losses, will benefit by taking action even when the forecast probability is low, while users with high $\alpha$ will obtain value by taking action only when the probability of the event is forecast to be high. An inappropriate choice of $p_t$ can reduce forecast value substantially. For example, a decision maker with $\alpha = 0.1$ will receive over 40% relative value by acting when the EPS probability is 10% or more, but will gain none at all by not acting until the forecast probability exceeds 50%.

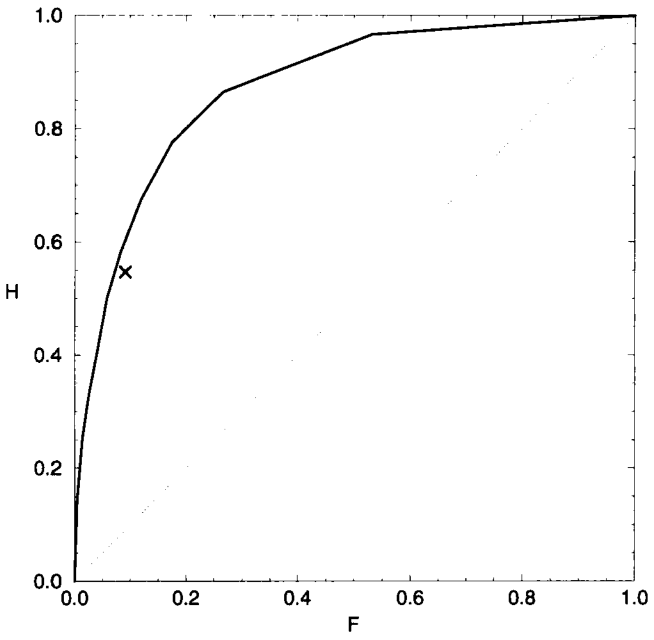

Figure 2. Relative operating characteristic (ROC) for Ensemble Prediction System (EPS) forecasts of 850 hPa temperature anomalies $T > +4$ K over Europe at day 6 for January and February 1998. Curve shows ROC for the EPS probability-forecast; the hit rate $H$ and false-alarm rate $F$ for the deterministic control-forecast are denoted by ×. The area under the ROC is $A = 0.875$.

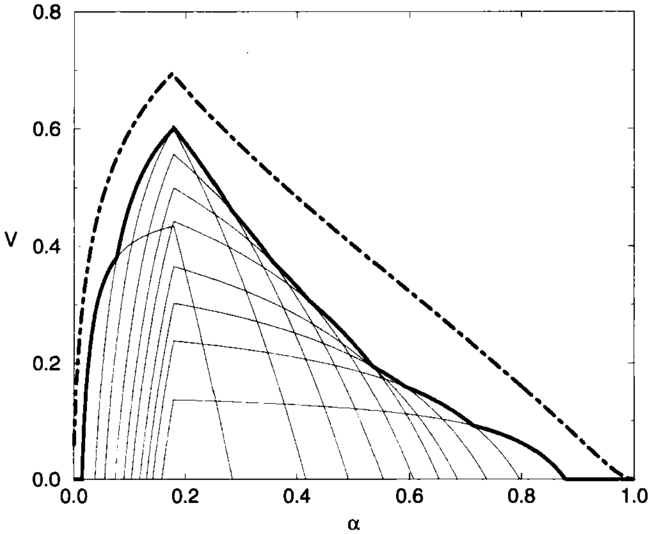

Figure 3. Relative value $V$ for Ensemble Prediction System (EPS) forecasts of 850 hPa temperature anomalies $T > +4$ K over Europe at day 6 for January and February 1998. Thin curves show $V$ for various probability-thresholds $p_t$; the envelope of these curves (heavy solid line) shows the overall relative value of the EPS, obtained by choosing the optimal value of $p_t$ for each value of the cost–loss ratio $\alpha$. The relative value of a 'perfect EPS' is also shown (dash-dotted line).

This example illustrates the importance of providing probability information to users: the value of the EPS forecasts depends significantly on the choice of $p_t$ and on the user's $\alpha$. There is no single threshold for which the EPS has value for all users—to benefit from the forecasts, different users must use different thresholds. If the EPS forecast is reduced to a single deterministic forecast for all users, for instance by using the ensemble mean or by choosing an arbitrary threshold, the value to some users will be reduced and may even be eliminated.

A probability-forecast system is said to be reliable if the event occurs on a proportion $p$ of those occasions where the forecast probability is $p$, i.e. the forecasts may be taken at face value (Stanski *et al.* 1989; Murphy 1993). For a reliable forecast-system, maximum value for a particular user will be obtained by choosing $p_t = \alpha$. More generally, an appropriate value of $p_t$ may be chosen to compensate for the forecast bias.

To provide maximum value to users with small $\alpha$, the EPS will need to resolve equally small probability-thresholds. The variation of value with $p_t$, and the implications for ensemble size are discussed in more detail in section 6.

Maximum value over all values of $\alpha$ occurs at $\alpha = \bar{o}$, as for the deterministic case, and is given by

$$V_{\max} = \max_{p_t}(H - F) = \max_{p_t}(KS) \equiv KS_{\max}. \qquad (14)$$

Thus for reliable forecasts, maximum relative value is $V_{\max} = KS(p_t = \bar{o})$.

We have defined $V$ as value relative to the value of a perfect deterministic forecast. In practice, the inherent uncertainty in our estimates of the current state of the atmosphere mean that perfect deterministic forecasts, and hence perfect relative value ($V = 1$), are unattainable. An estimate of the potential relative value of the EPS, given current analysis uncertainty, can be made by considering the ensemble as a 'perfect EPS': instead of the EPS being verified against the real outcome, the distribution of solutions in the EPS is assumed to be an accurate representation of the potential outcomes. This effectively eliminates model errors. The perfect EPS can be evaluated by using one ensemble member as the verification rather than the real analysis (Buizza and Palmer 1998). Alternatively, for the point probability-forecasts studied in the present work, scores for the perfect EPS may be derived directly from the forecast probabilities. Let the function $g(p)$ denote the frequency with which the EPS forecasts a given event with probability $p$. Let $p'(p)$ be the frequency with which the event occurs when the forecast probability is $p$. These two functions completely specify the performance of the EPS in forecasting the event. For example, hit rate $H_{p_t}$ can be written as

$$H_{p_t} = \left[ \left\{ \int_{p_t}^{1} p'(p) g(p) \, \mathrm{d}p \right\} \Big/ \left\{ \int_{0}^{1} p'(p) g(p) \, \mathrm{d}p \right\} \right]. \qquad (15)$$

For a 'perfect EPS', $p'(p) = p$ for all values of $p$, and the skill and value of the ensemble depends only on the distribution of forecast probabilities $g(p)$. This method is statistically more robust than evaluating against a randomly chosen ensemble member. (For a large enough ensemble and sample size the two methods will give the same result.)

The relative value of the perfect EPS, $V_{\text{perf}}$, is shown in Fig. 3. The relative value of the real EPS forecasts is about 0.1 below that of the perfect EPS for almost all values of $\alpha$, with greater potential gains when $\alpha$ is small. Just as the difference between the relative value $V$ of the EPS and $V_{\text{perf}}$ gives an indication of the potential for improvement by elimination of model errors, the difference between $V_{\text{perf}}$, and the perfect *deterministic* limit $V = 1$ indicates the importance of uncertainty of the initial condition in limiting the value which can be expected in practice. In the perfect EPS context, $V_{\mathrm{perf}}$ can be improved only by reducing the initial uncertainty, i.e. by reducing analysis errors. Future developments in data assimilation are designed to lead to better analyses and lower initial uncertainties—this will lead to higher relative value.

## 5. COMPARISON OF DETERMINISTIC AND PROBABILISTIC FORECASTS

One of the benefits of the ROC is that it allows direct comparison of deterministic and probabilistic forecast systems. $H$ and $F$ for the control forecast (Table 3) are plotted together with the EPS probability ROC in Fig. 2 for +4 K. The point for the control forecast lies below the EPS ROC; since for the same $F$ a greater hit rate is obtained using the EPS probabilities, this implies that the control forecasts are less useful than the EPS forecasts. This will be true for all users, irrespective of $\overline{o}$ and $\alpha$, since, if these are fixed and $F$ are fixed, $V$ increases with $H$ (Eq. (8)). Although the difference between the control point and the EPS ROC in Fig. 2 is small and not necessarily statistically significant, comparison of the corresponding value-curves highlights the advantage of the probability forecasts over deterministic forecasts of similar quality (Fig. 4). The substantial additional skill of the EPS for many users is a result of their being able to choose the appropriate value of $p_t$. This flexibility greatly increases the range of users who will benefit from the EPS forecasts. The relative value of a perfect EPS is also shown; the improvement over the real EPS is similar for each event, increasing $V$ by 0.1–0.2 for most users.

The three points (0, 0) ($H_{\mathrm{control}}$, $F_{\mathrm{control}}$) and (1, 1) may be considered to define a ROC curve for the deterministic forecast. As for the probability forecast, the accuracy of this forecast can then be measured by the area $A$ under the curve (i.e. the area of the quadrilateral defined by these points and (1, 0)). It is straightforward to show that, for the deterministic forecast, the area skill $ASS$ (Eq. (13)) equals the standard $KS$ skill score:

$$KS = 2A - 1 \equiv ASS. \tag{16}$$

The skill of the deterministic control-forecasts and of the EPS probability-forecasts of the +4 K anomaly are compared for forecast days 3 to 10 (D3 to D10) in Table 4. For the EPS, $ASS$ is substantially higher than $KS_{\mathrm{max}}$ (Eq. (14)), giving the impression of greater skill and a larger improvement over the control forecast. A deterministic forecast with skill equal to $KS_{\mathrm{max}}$ will still have less value than the probability forecasts for a range of $\alpha$ because of the flexibility of varying $p_t$ for different users. The difference between $ASS$ and $KS_{\mathrm{max}}$ for the ensemble reflects this additional benefit, although $ASS$ is difficult to interpret quantitatively. In the following, let us use $KS_{\mathrm{max}}$ rather than $ASS$ as an index of EPS skill because of its simple interpretation as the maximum value of the ensemble skill. Use of $ASS$ instead would show a greater skill advantage for the EPS in the following analysis.

Comparison of the $KS$ scores for EPS and control show the probability forecasts have a substantial advantage at all lead times. Indeed, the D10 probability-forecast is as skilful (has the same maximum value) as the D6 control. This gives a measure of the benefit of the EPS—for the single deterministic forecast to have the same skill as the current EPS at D10, the forecast model would need to be improved until deterministic D10 forecasts were as skilful as the present D6 forecasts. This skill advantage is shown for all temperature events in Table 4 (missing values indicate where the EPS was more skilful than the D3 control; scores are not available for D1 and D2 forecasts). For most

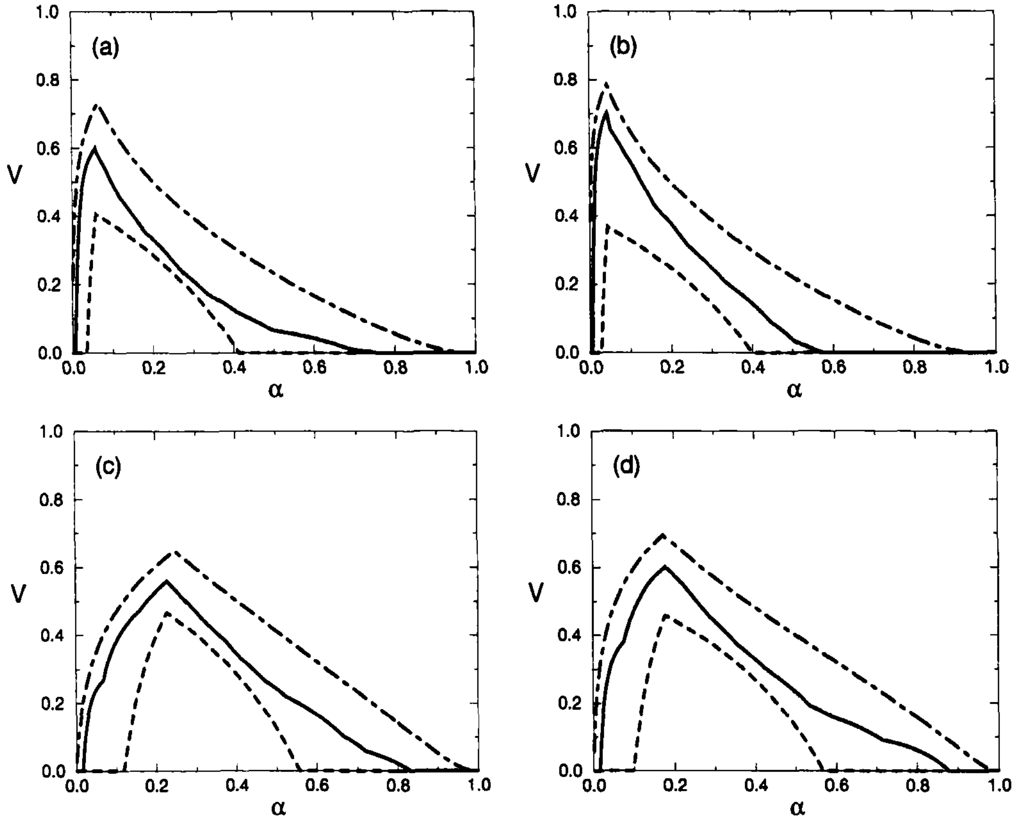

Figure 4. Relative values $V$ plotted against cost–loss ratio $\alpha$ for four 850 hPa temperature anomaly events over Europe at day 6 for January and February 1998: (a) $T < -8$ K; (b) $T > +8$ K; (c) $T < -4$ K; (d) $T > +4$ K. Lines show deterministic control-forecasts (pecked line); Ensemble Prediction System (EPS) probability-forecasts (solid line) and 'perfect EPS' (dash-dotted line).

TABLE 4. SKILL OF CONTROL FORECAST AND EPS PROBABILITY-FORECASTS FOR 850 hPa TEMPERATURE ANOMALIES OVER EUROPE FOR FORECAST DAYS 3–10, FOR JANUARY AND FEBRUARY 1998*.

| | | Forecast day | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Score | Event | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| EPS *ASS* | $T > +4$ K | 0.953 | 0.933 | 0.909 | 0.875 | 0.848 | 0.830 | 0.810 | 0.792 |
| EPS $KS_{\max}$ | $T > +4$ K | 0.794 | 0.737 | 0.674 | 0.602 | 0.531 | 0.509 | 0.480 | 0.454 |
| Control *KS* | $T > +4$ K | 0.708 | 0.625 | 0.518 | 0.457 | 0.406 | 0.355 | 0.303 | 0.274 |
| $D_{\text{adv}}$ | $T > +4$ K | | | 1.59 | 1.79 | 2.12 | 2.85 | 3.38 | 3.94 |
| $D_{\text{adv}}$ | $T > +8$ K | | | | | 3.61 | 4.38 | 4.91 | 5.56 |
| $D_{\text{adv}}$ | $T < -4$ K | | 0.50 | 0.54 | 0.77 | 1.20 | 1.54 | 1.95 | 2.49 |
| $D_{\text{adv}}$ | $T < -8$ K | | | | 2.15 | 2.32 | 2.78 | 3.22 | 3.96 |

*See text for details.

lead-times and events studied here, the EPS has an advantage ($D_{\text{adv}}$) of 2–4 days over the control. This is a considerable difference and highlights the importance of the EPS—improvements in deterministic skill of such magnitude are generally achieved only over many years of development of all aspects of model and data-assimilation formulation (Simmons *et al.* 1995).

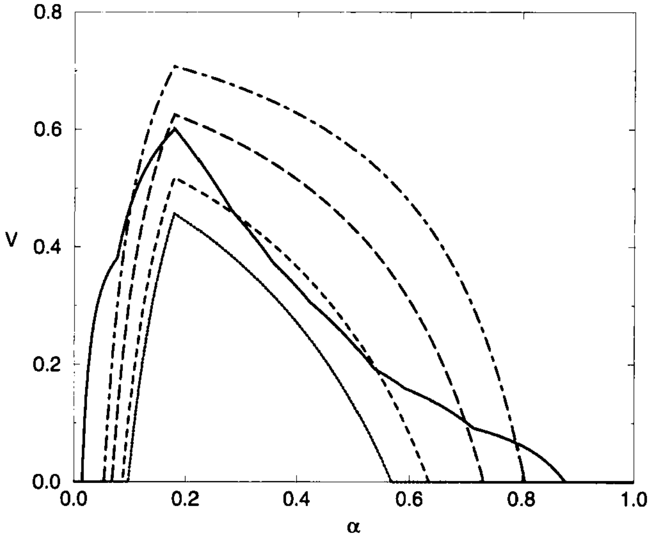

Figure 5. Variation with cost–loss ratio $\alpha$ of the relative value $V$ of Ensemble Prediction System (EPS) probability-forecasts at day 6 (solid line) and deterministic control-forecasts for different lead-times (day 6, dotted; day 5, pecked; day 4, long dashed; day 3, dash-dotted) for 850 hPa temperature anomalies $T > +4$ K over Europe at day 6 for January and February 1998.

TABLE 5. SKILL AND MAXIMUM INCREASE IN VALUE OF EPS D10 OVER CONTROL D3 FORECASTS OF TEMPERATURE OVER EUROPE

| | January and February 1998 | | |
|---|---|---|---|
| Event | *KS*(D3 control) | $KS_{\max}$(D10 EPS) | Max{*V*(D10 EPS) − *V*(D3 control)} |
| $T < -8$ K | 0.688 | 0.404 | 0.194 |
| $T < -4$ K | 0.728 | 0.343 | 0.070 |
| $T > +4$ K | 0.708 | 0.454 | 0.110 |
| $T > +8$ K | 0.648 | 0.514 | 0.328 |

The analysis of the previous paragraph demonstrated the advantage of the EPS in terms of skill or, equivalently, maximum relative value. For different users, the relative value of EPS and the control forecasts varies substantially. Figure 5 shows how lead time changes value for all users for $T > +4$ K. One forecast-system is said to be sufficient for another if it provides greater value for all users (Ehrendorfer and Murphy 1988). It is clear from Fig. 5 that even the D3 control forecast is not sufficient for the D6 EPS. Table 5 illustrates the sufficiency of the D3 control for the D10 EPS for the four temperature-thresholds. For each event, the D3 control skill (maximum value) is, as expected, substantially greater than the D10 EPS skill. However, for every event there are some users for whom the D10 EPS has considerably more value than the D3 control. The D3 control forecast is not sufficient for the D10 EPS for any of the temperature events.

It is users with low $\alpha$ who derive the greatest benefit from the increase in value of the EPS over the shorter-range control-forecast (Fig. 5). This has particular significance when the potential savings from more timely warnings are taken into account. Users with more time to prepare for protective action may be able to reduce the cost of taking it and so reduce $\alpha$ towards the region where the EPS has the greatest advantage.

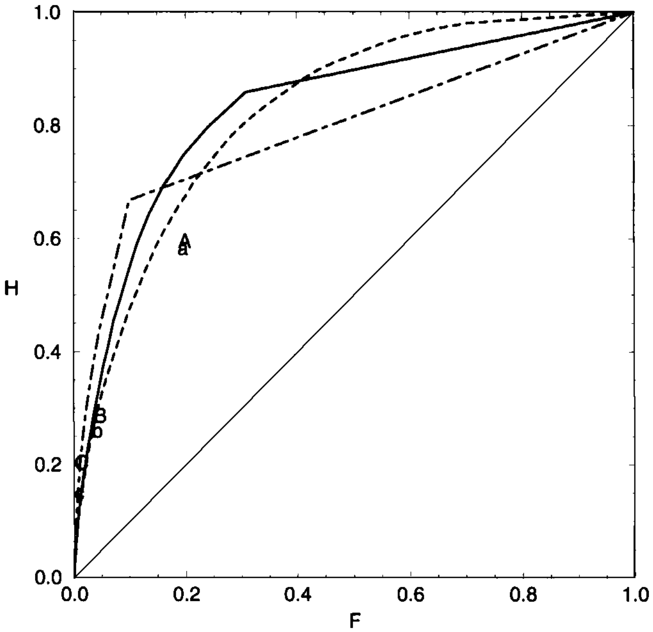

Figure 6. Relative operating characteristics (ROCs) for Ensemble Prediction System (EPS) forecasts of 12-hour total precipitation exceeding 1 mm (pecked line), 5 mm (solid) and 10 mm (dash-dotted) over Europe at day 5 for winter 1996/97. Curves are for EPS probability-forecasts calculated using all possible probability-thresholds (approximately at 2% intervals). Hit rates $H$ and false-alarm rates $F$ for the deterministic EPS control-forecast (lower-case letters) and high-resolution operational forecast (upper case) are also shown (a,A, 1 mm; b,B, 5 mm; c,C, 10 mm).

## 6. The effect of ensemble size on forecast value

It was shown in section 4 that for a given user, maximum relative value is gained when $p_t$ is equal to the user's $\alpha$. It is, therefore, important to consider the variation of $V$ with $p_t$. If small changes in $p_t$ lead to significant differences in value, then the EPS resolution needs to be sufficient to resolve the required probability-thresholds.

The relationships between ensemble size, $p_t$ and $V$ are illustrated using EPS forecasts of precipitation over Europe in the 12-hour period $t + 108$ to $t + 120$ (D4.5 to D5) exceeding a given threshold. Three events are considered which represent different climatological frequencies: total precipitation exceeding 1 mm, 5 mm and 10 mm. The EPS forecasts are verified against the $t + 12$ to $t + 24$ accumulated-precipitation forecasts from the operational deterministic high-resolution (T213) ECMWF model; results are for 92 cases from winter 1996–97 (Buizza *et al.* (1999); see their appendix for a detailed discussion of this choice of verification data).

In addition to ensemble size, an important factor in determining an optimal ensemble-configuration is the resolution of the forecast model. For a given allocation of computing power, a compromise must be reached between ensemble size and model resolution (Buizza *et al.* 1998). Precipitation forecasts are perhaps particularly sensitive to resolution. For this reason, the operational deterministic high-resolution (T213) forecast is assessed in addition to the EPS control forecast to give an indication of the effect of the reduced EPS-resolution.

ROCs for the three events are shown in Fig. 6; the area under each curve and the maximum relative value of the system are given in Table 6. The EPS has skill for all events and the forecasts all have substantial relative value.

TABLE 6. EPS PERFORMANCE FOR PRECIPITATION EVENTS

| Precipitation threshold | $\bar{o}$ | ROC area | $V_{\max}$ | $KS_{\text{control}}$ | $KS_{\text{operational}}$ |
|---|---|---|---|---|---|
| 1 mm | 0.289 | 0.826 | 0.50 | 0.338 | 0.397 |
| 5 mm | 0.0552 | 0.831 | 0.56 | 0.218 | 0.239 |
| 10 mm | 0.0135 | 0.794 | 0.57 | 0.140 | 0.190 |

For each precipitation amount, the ROC points for both deterministic forecasts lie below the corresponding EPS-probability ROC and the maximum value (or skill, *KS*) is substantially lower than that of the EPS. The skill of the high-resolution operational forecast is, however, greater than that of the lower-resolution EPS control, indicating the potential for improvement to the EPS by using a higher-resolution forecast-model. Although the overall value of the EPS is greater than the single high-resolution forecast, there are circumstances where the lower-resolution EPS fails to capture an important extreme event, whereas an integration run at enhanced resolution provides a successful forecast (A. Hollingsworth, personal communication). In such circumstances, while a single high-resolution forecast will not always be correct, it can provide valuable additional information to forecasters on the potential effect of resolution.

For the EPS, $V_{\max}$ is higher for the 5 mm and 10 mm events than for the 1 mm threshold, but skill (as determined by ROC area) is lower for the 10 mm event than for the other precipitation-thresholds. It is apparent from Fig. 6 that whereas points on the ROC are reasonably evenly spaced for precipitation over 1 mm, data for the ROCs for 5 mm and 10 mm are increasingly restricted towards the lower-left portion of the diagram. If attention is focused on just those parts of the ROC curves where data are available for all three precipitation amounts, it is clear that the EPS performance is better for heavier precipitation. (For a given false-alarm rate $F$, the highest hit-rate $H$ is achieved for the 10 mm event, and the lowest for the 1 mm event).

The reason for the restricted coverage of the larger-event ROCs is related to the lower observed frequency of the heavier-rainfall events (Table 6) and the values of $p_t$ used to calculate the values of $H$ and $F$. If the observed frequency of the event is low relative to the forecast $p_t$, and the forecast system is reasonably reliable (i.e. rare events are forecast rarely), the great majority of points in the contingency table (Table 2) for this $p_t$ will be in the No/No cell (cell '*a*' in Table 2); hence $F$ will be low and $H$ will be limited accordingly. To extend the ROC data-points towards the top right of the diagram, additional lower-probability thresholds must be considered. In practice, the resolution of probability thresholds is limited by the size of the ensemble. For an $N$-member ensemble, all possible hit and false-alarm rates can be calculated by taking $p_t$ at intervals $\mathrm{d}p_t = 1/N$. The ROCs in Fig. 6 were calculated using all values of $p_t$ available from the 51-member EPS, i.e. $p_t$ intervals of approximately 0.02. Although this resolution is sufficient for the relatively frequent 1-mm precipitation event, a larger ensemble would be needed to specify the full ROC for the rarer 10-mm event.

The ROC is often generated using a fixed set of probability thresholds at 10% intervals (Stanski *et al.* 1989; Buizza and Palmer 1998). This use of a limited set of probability thresholds may give a misleading impression of the potential performance of the forecast system. The effect is illustrated in Fig. 7 and Table 7 in which the ROCs and scores for each precipitation event have been recalculated using $\mathrm{d}p_t = 0.1$, rather than the full set of probability thresholds used for Fig. 6. There is little effect on the ROC for the relatively frequent 1-mm event; the area decreases slightly but $V_{\max}$ is unchanged. However, the scores for the larger precipitation amounts are substantially

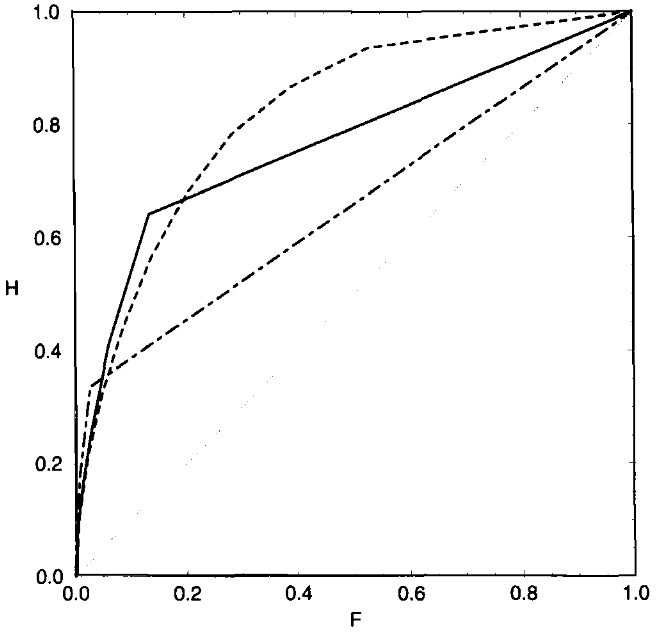

Figure 7. As Fig. 6, but for ROC curves calculated using probability thresholds $p_t$ at 10 % intervals.

TABLE 7. EPS PERFORMANCE FOR PRECIPITATION EVENTS ($dp_t$ = 0.1)

| | 51 members | | 10 members | |
|---|---|---|---|---|
| Precipitation threshold | ROC area | $V_{max}$ | ROC area | $V_{max}$ |
| 1 mm | 0.819 | 0.50 | 0.800 | 0.47 |
| 5 mm | 0.764 | 0.51 | 0.711 | 0.40 |
| 10 mm | 0.656 | 0.31 | 0.638 | 0.28 |

degraded and give the misleading impression that the 10-mm event is considerably less skilful than the 1-mm event.

The variation of relative value with $\alpha$ is shown for the two sets of probability thresholds in Fig. 8. For all precipitation events there is little effect from changing $dp_t$ for $\alpha > 0.1$. However, large differences are apparent for smaller values of $\alpha$. The curves are plotted with a logarithmic $x$-axis so that this can be seen clearly. Although $V_{max}$ does not increase for $dp_t = 0.02$ for the 1-mm event, there are still substantial gains for users with low $\alpha$: use of the smaller $dp_t$ gives values of $V$ up to 20% where the $dp_t = 0.1$ probability-thresholds gave no value. The differences in value are notably larger for the more extreme events. Also shown in Fig. 8 are estimates of the relative value which could be obtained with a sufficiently large ensemble. These estimates were derived using a parametrized model of the ROCs of Fig. 6 (Mason 1982; Harvey *et al.* 1992); the method is summarized briefly in an appendix. There is potential for substantial improvement for users with low $\alpha$, particularly for the more extreme events (as $\bar{o}$ decreases, so too does the probability threshold $p_t$ for maximum value).

For values of $\alpha$ larger than about 0.1, there seems to be little benefit to be obtained solely from increasing the resolution of the probability threshold. However, increasing

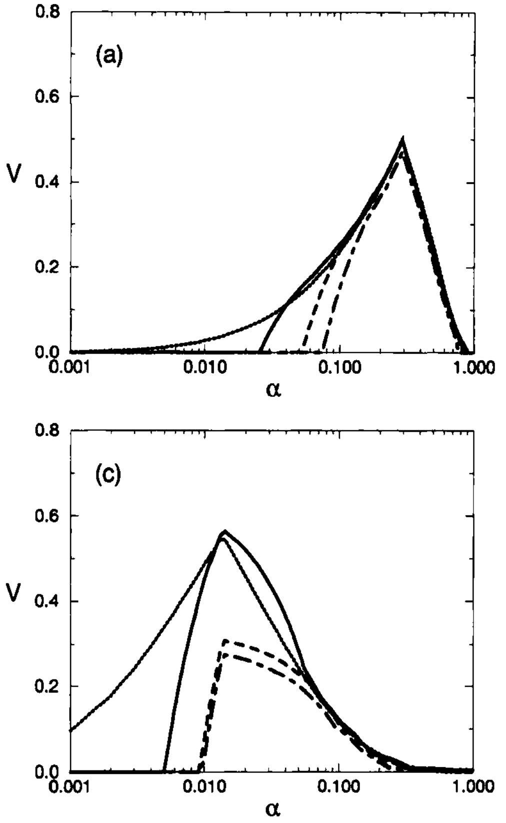

Figure 8. Variation with cost–loss ratio $\alpha$ of the relative value $V$ of Ensemble Prediction System (EPS) probability-forecasts of 12-hour total precipitation exceeding stated amounts over Europe at day 5 for winter 1996/97: (a) 1 mm; (b) 5 mm; (c) 10 mm. $V$ is shown for different resolutions of probability threshold $p_t$: standard 10% intervals, $\mathrm{d}p_t = 0.1$ (pecked line); all possible thresholds, $\mathrm{d}p_t = 0.02$ (solid line); estimate of potential relative value as $\mathrm{d}p_t \rightarrow 0$ (dotted line). Also shown is relative value $V$ calculated using ten ensemble-members for $\mathrm{d}p_t = 0.1$ (dash-dotted line).

ensemble size may be of value in providing more reliable estimates of forecast probability. The effect of ensemble size is examined by using just 10 of the 51 ensemble members to calculate $H$ and $F$ for the standard $p_t$ thresholds ($\mathrm{d}p_t = 0.1$). ROC area and maximum value are both reduced (Table 7), and relative value is lower for all users than for 51 members (Fig. 8).

These results indicate that increasing ensemble size will bring benefits, both by allowing finer resolution of probability thresholds and by giving improved estimates of the forecast-probability distribution. The potential-value estimates of Fig. 8, which only consider the effect of increasing resolution of $\mathrm{d}p_t$, should, therefore, be considered as lower bounds to the improvement obtainable with larger ensembles.

## 7. Conclusions

The EPS is an important component of the operational forecasting capability at ECMWF, complementing the operational deterministic forecast with probabilistic information reflecting the uncertainty in the analysed atmospheric state. A simple decision-analytic model has been used to study the potential economic value of EPS forecasts of temperature and precipitation.

The EPS has considerable value throughout the medium range, although different users will benefit to a greater or lesser extent, depending on their specific economic costs. Comparison of the EPS probability-forecasts with deterministic predictions from the control integration demonstrates the advantage of the probabilistic approach (Murphy 1993). Probability forecasts are generally more useful than deterministic forecasts of comparable quality because the user can select a probability threshold appropriate to his needs. A forecaster's arbitrary prescription of such a threshold without knowledge of a particular user's requirements can severely reduce the value of the system to that user.

The value of the operational EPS was compared to that of a 'perfect EPS' in which model errors are eliminated. The potential improvement in relative value $V$ is 0.1–0.2 for most users for the four temperature events considered. The substantial difference between the relative value of the perfect EPS and the perfect deterministic limit ($V = 1$) is an indication of the effect of the initial uncertainty in limiting the value which can be expected in practice. If the initial uncertainty can be reduced by improvements to the forecast model and data-assimilation system, this will lead to improvements in the potential value of the EPS.

In our point verification, the EPS probability-forecasts have several days' advantage over the corresponding control forecast. In other words, the additional information in the EPS, reflecting the uncertainty in the initial conditions, provides a benefit to users equivalent to many years' development of the deterministic forecast-model and assimilation-system. In fact, for some users, EPS probability-forecasts for 10 days ahead have more value than day-3 deterministic forecasts.

The analysis presented in the present paper demonstrates the overall benefit of the EPS probability-forecasts over a single deterministic forecast. For precipitation forecasts, which may be particularly sensitive to model resolution, the EPS was shown to be superior to the higher-resolution T213 deterministic forecast. However, the improvement of the T213 forecast over the $\mathrm{T_L}159$ EPS control indicates the potential benefit of increased resolution. Despite the overall benefit of the EPS, there are circumstances in which the EPS resolution is too low to capture an important event. Whereas a single high-resolution forecast will not always be correct, it can provide valuable additional information on the sensitivity of a particular situation to model resolution; it is for this reason that a high-resolution deterministic forecast is run operationally at ECMWF alongside the lower-resolution EPS.

Maximum relative value occurs when the cost–loss ratio $\alpha$ is equal to the observed frequency of the event. For a particular user, the greatest value of a reliable probability-forecast is found for threshold probability $p_{\mathrm{t}}$ equal to the user's $\alpha$. Therefore, users with small $\alpha$ will receive maximum benefit only if equally small probability-thresholds can be resolved by the EPS. Recent studies of ensemble size (Buizza and Palmer 1998; Buizza *et al.* 1998, 1999) have followed the common practice of using a standard set of probability thresholds (at 10% intervals). While some benefit was found for all users by using all 51 ensemble-members rather than just 10 to estimate these thresholds, a far greater increase in value for small $\alpha$ was found by using the full ensemble to discriminate smaller probability-thresholds. Further increases in ensemble size will give significant additional improvement for low $\alpha$, especially for more extreme events.

Ideally, the weather sensitivity, decision processes and relevant costs of all potential users would be known. It would then be straightforward to calculate the overall value of the EPS. Proposed changes to the forecast system could be evaluated in terms of net benefit, and the effect on any given user could be determined. In practice, little is known about the decision-making processes of many weather-sensitive activities (users themselves are often unclear about this information). One possible approach (Roebber and Bosart 1996) is to study overall value as a function of various hypothetical distributions for $\alpha$, although the absolute costs of different users and the relative importance of different weather-events will also have a significant impact on total value.

However, until more is known about the spectrum of potential users, it is important to be aware of the effect of the EPS on the full range of $\alpha$. As Roebber and Bosart (1996) pointed out, high values of $\alpha$ are difficult for a business to sustain; also, competition between businesses is likely to act to reduce $\alpha$. The provision of forecasts with longer lead-time may also serve to reduce costs (see section 5). Finally, for extreme events, potential losses may, in many instances, greatly exceed the cost of protection. The advantages of the EPS, and potential benefits of increased ensemble size for low $\alpha$, may thus be of considerable significance.

Although the cost and losses for a particular user may be difficult to determine, the present value-study offers forecast verification in a form relevant to the user's needs. The decision framework, although simple, shows that users will not all feel the same benefits of the EPS, nor will they be affected equally by changes to the forecasting system.

## ACKNOWLEDGEMENTS

I thank Tim Palmer, Adrian Simmons, Tony Hollingsworth and two anonymous referees for their helpful comments on an earlier version of this paper.

## APPENDIX

### *The Gaussian model for the relative operating characteristic*

The relative operating characteristic (ROC) curves presented in the present paper have all been produced empirically as plots of hit rate $H$ and false-alarm rate $F$ derived from forecast-verification data. Mason (1982) and Harvey *et al.* (1992) demonstrate that a parametrized model of the ROC, derived from signal-detection theory (SDT) provides a good fit to such empirical-forecast data.

The model assumes that information about the occurrence (or not) of an event can be represented by a single one-dimensional variable $X$. Uncertainty about the outcome is reflected in case-to-case variations in $X$, which are given by two fixed probability-distributions, one for the distribution of $X$ given the event occurs, $\phi_s(X)$ (the signal distribution in SDT), and another for the distribution of $X$ given the event does not occur, $\phi_n(X)$ (the noise distribution).

For any given value of $X_c$ of $X$ (the decision criterion), $H$ and $F$ are then simply the areas to the right of $X_c$, under the respective probability density functions (pdfs):

$$H = P(X > X_c \mid \text{event occurs}) = \int_{X_c}^{\infty} \phi_s(x)\,\mathrm{d}x \tag{A.1}$$

$$F = P(X > X_c \mid \text{event does not occur}) = \int_{X_c}^{\infty} \phi_n(x)\,\mathrm{d}x. \tag{A.2}$$

So, as the decision criterion $X_c$ varies, sequences of values of $H$ and $F$ are produced which trace out the ROC for the system. The model and the ROC are thus defined by the distributions $\phi_n(X)$ and $\phi_s(X)$. The simple case of both distributions being Gaussian is generally found to produce good results (Mason 1982; Harvey *et al.* 1992).

If both distributions are Gaussian (means $\mu_s$, $\mu_n$; standard deviations $\sigma_s$, $\sigma_n$) then $H$ and $F$ can be expressed as areas under the standard Gaussian distribution

$$H = \int_{X_c}^{\infty} \phi_s(x)\,\mathrm{d}x = \int_{Z_s}^{\infty} \phi(z)\,\mathrm{d}z \tag{A.3}$$

where

$$z_s = (X_c - \mu_s)/\sigma_s, \tag{A.4}$$

with a similar expression for $F$. The quantities $z_s$ and $z_n$ are related by

$$z_s = (\sigma_n z_n + \mu_n - \mu_s)/\sigma_s. \tag{A.5}$$

Thus, a set of values of $H$ and $F$ may be transformed to the corresponding standardized Gaussian deviates $z_s$ and $z_n$ (Eq. (A.3)); the strength of the linear relationship between $z_s$ and $z_n$ (Eq. (A.5)) is a measure of the validity of the Gaussian model for the original data. Correlation between the transformed variables is greater than 0.99 for all the events discussed in the present paper (statistically significant beyond the 1% level), confirming that the Gaussian model is applicable to the EPS data.

The Gaussian model is completely specified by two parameters, usually chosen as the distance between the means of the two Gaussian distributions (normalized by the standard deviation of $\phi_n$) and the ratio of the two standard deviations. These may be estimated from a set of empirical ROC data. The model may then be used to construct a parametrized version of the original ROC, but with any desired resolution of probability threshold. In this way, the model may be used to provide an estimate of the potential benefit of having finer resolution in $\mathrm{d}p_t$.

## REFERENCES

| | | |
|---|---|---|
| Buizza, R., Hollingsworth, A., Lalaurette, E. and Ghelli, A. | 1999 | Probability precipitation prediction using the ECMWF Ensemble Prediction System. *Weather and Forecasting*, **14**, 168–189 |
| Buizza, R. and Palmer, T. N. | 1998 | Impact of ensemble size on ensemble prediction. *Mon. Weather Rev.*, **126**, 2503–2518 |
| Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, L., Hamrud, M., Hollingsworth, A., Simmons, A. and Wedi, N. | 1998 | Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **124**, 1935–1960 |
| Ehrendorfer, M. and Murphy, A. H. | 1988 | Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy. *Mon. Weather Rev.*, **116**, 1757–1770 |
| Gandin, L. S. and Murphy, A. H. | 1992 | Equitable skill scores for categorical forecasts. *Mon. Weather Rev.*, **120**, 361–370 |
| Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M. and Mross, E. F. | 1992 | The application of signal detection theory to weather forecasting behaviour. *Mon. Weather Rev.*, **120**, 863–883 |
| Katz, R. W. and Murphy, A. H. (Eds.) | 1997a | *Economic value of weather and climate forecasts.* Cambridge University Press, UK |
| | 1997b | 'Forecast value: prototype decision-making models'. In *Economic value of weather and climate forecasts.* Cambridge University Press, UK |
| Mason, I. | 1982 | A model for assessment of weather forecasts. *Aust. Meteorol. Mag.*, **30**, 291–303 |
| Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. | 1996 | The ECMWF Ensemble Prediction System: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73–119 |
| Murphy, A. H. | 1977 | The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Mon. Weather Rev.*, **105**, 803–816 |
| | 1993 | What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–293 |

| | | |
|---|---|---|
| Murphy, A. H. | 1994 | Assessing the economic value of weather forecasts: an overview of methods, results and issues. *Meteorol. Apps.*, **1,** 69–73 |
| Murphy, A. H. and Ehrendorfer, M. | 1987 | On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Weather and Forecasting*, **2,** 243–251 |
| Palmer, T. N., Molteni, F., Mureau, R. and Buizza, R. | 1993 | 'Ensemble prediction', in ECMWF Seminar Proceedings *Validation of models over Europe: Vol. I*, ECMWF, Shinfield Park, Reading, RG2 9AX, UK |
| Palmer, T. N., Brankovic, C. and Richardson, D. S. | 2000 | A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.*, in press |
| Roebber, P. J. and Bosart, L. E. | 1996 | The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather and Forecasting*, **11,** 544–559 |
| Simmons, A. J., Mureau, R. and Petroliagis, T. | 1995 | Error growth and estimates of predictability from the ECMWF forecasting system. *Q. J. R. Meteorol. Soc.*, **121,** 1739–1771 |
| Stanski, H. R., Wilson, L. J. and Burrows, W. R. | 1989 | 'Survey of common verification methods in meteorology'. World Weather Watch Technical Report No. 8, WMO/TD. No. 358, World Meteorological Organization, Geneva, Switzerland |
| Wilks, D. S. | 1995 | *Statistical methods in the atmospheric sciences*. Academic Press, London, UK |
| Wilks, D. S. and Hamill, T. M. | 1995 | Potential economic value of ensemble forecasts. *Mon. Weather Rev.*, **125,** 3565–3575 |